IP Based Distributed Virtual SAN

By
Sheng (Ted) Tai Tsao
8/1/2002

Field of the Invention

This invention is the continuation of the previous invention, application number 60/401,238, of "Concurrent Web Based Multi-Task Support for Control Management System", which focus on web based multi-tasking support for web-console in the central controlled distributed scalable virtual machine environment. The present invention focus on distributed IP SAN based storage service and other distributed services in the central controlled distributed scalable virtual machine environment. It relates generally to IP based out-band accessed distributed virtual SAN infrastructure, its automatic configuration, its storage volumes allocation and accessing services. This invention also presents the applicability of the principles of IP based distributed virtual SAN service to other services and applications in a similar environment.

Background Information

a) Terminology:

CCDSVM:

It is an abbreviation for central controlled distributed scalable virtual machine system. The CCDSVM allows a control management station to control a group of systems and provide distributed services to client system in Intranet and Internet as well as in LAN environment.

Storage Media:

Storage media includes magnetic hard disk drives, solid state disk, optical storage drive, and memory card etc.

Storage Connection and Control Media:

Storage connection and control media includes controller of IDE, SCSI, Fibre optical, Ethernet, USB, or may be wireless media and all related cables etc. Each controller of storage media such as Raid, IDE, or SCSI controller may control multiple storage media drives on a system.

Storage System:

Storage system includes one or more storage media and the storage connection and control media. Storage system also contains related software modules to deliver storage services.

SAN:

SAN stands for storage area of network. It is a storage system, which provides computer host with block data service through storage connection media, such as Fibre-optical cable, Ethernet cable or other media by using protocol based on IP, non-IP based such as Fibre-Channel, or others. IP SAN uses IP based protocol to provide storage raw block data services. All discussion of SAN in this invention are within the scope of a model of central controlled distributed scalable virtual machine (CCDSVM).

DNS:

It stands for domain name server of network technology, which is a Internet software infrastructure. It helps any system on the net to find its peer target system's network address in order to send the message to its peer system.

SNMP:

An abbreviation for "Simple Network Management Protocol", which is a standard Internet protocol. The SNMP trap is a UDP packet sent

by SNMP daemon on a SNMP agent system to SNMP network management station through network link.

b) Figures:

1) Distributed Virtual SAN Infrastructure.
2) The Actual Components of Distributed Virtual SAN.
3) Virtual SAN Automatic Configuration Protocol.
4) Virtual SAN Auto Configuration Protocol Packet format.
5) Example of Storage Volume Information of an IP SAN Unit.
6) A Hypothetical Example of Storage Volume Requests and Assignment.
7) Direct Attached Storage System.
8) In-Bound Accessed Virtual Storage System.
9) A Simplified Diagram of Central Controlled Distributed Scalable Virtual Machine System.
10) A Simplified Diagram of Disaster Recovery Scheme of Distributed Virtual SAN Infrastructure.

In the drawing, like elements are designed by like reference numbers.

Brief Description of the Invention

Today's corporate IT professional faces many challenges to handle the ever increasing information and data. This often requires many organizations to expand their storage capacity, manage storage systems and to keep the normal business run. Currently, The IP based NAS (network attached storage) effectively provides storage and services for end user's file system needs. On the other hand, at the enterprise level, the majority storage systems are still server directly attached (Fig. 7) and being accessed as raw block data devices through either traditional IDE, SCSI, Fibre Channel, or may be Ethernet.

The server direct attached storage system (Fog. 7) has many drawbacks, which are described as follow:

a) Currently, the most advance storage management system only capable to handle 4TB of data, which is far from good enough for enterprise storage management requirement.
b) The most of server directly attached storage has problems to expand its capacity. In some case, it is quite often to require purchasing a new server in order to expand the storage. In other cases, it also requires to shutdown the server and to stop the normal operation in order to expand the storage capacity.
c) The storage being attached can only be accessed by the attached server and can not be shared by other even a server's storage has spare capacity left while other server are in shortage of the storage capacity within a department or cross department in a organization.
d) Each attached storage system has to be managed separately and this is a nightmare for IT professional.
e) With the attached storage system, the backup/restore has to go through the data network, this will tax the data network performance.
f) The SCSI only allow 12 meter distance for data accessing with 15 storage devices while Fibre Channel is also limited to 10 kilometers long. This effectively prevents them from being the best choice for disaster recovery of the storage system.
g) The Fibre Channel based storage system cannot handle well for the interoperability. Also, Fibre Channel based storage system is expensive to build and to maintain.

There is a type of virtual SAN, which is in-band controlled and accessed (Fig. 8), with which the data path from hosts (1 of Fig. 8) to the SAN units (4 of Fig. 8) goes through virtual SAN control management station (2 of Fig. 8). It is not efficient in term of accessing the data by the hosts due to the virtual SAN control management

station can easily be a performance bottleneck. By same reason, the scalability of this type virtual SAN also is poor.

With the rapid development of high speed communication technology, the problems mentioned above can be solved by an IP based out-band accessed distributed virtual SAN infrastructure (Fig. 1) of this invention. With this invention, each hosts (1 of fig. 1) can directly access IP based SAN units (4 of Fig. 1) without go through control management station (3 of Fig. 1). The IP based out-band accessed distributed virtual SAN infrastructure (Fig. 1) actually represents an example of central controlled distributed scalable virtual machine system (CCDSVM) (Fig. 9). Wherein, each system units actually is a SAN unit (4 of Fig. 1), specifically is an IP based SAN unit.

With this invention, each SAN unit (4 of Fig.1) can be accessed by one or more hosts (1 of Fig.1) and each hosts can access one or more SAN units (Fig. 6). In addition, the storage accessing goes directly through communication link (2 of Fig. 1) between hosts (1 of Fig. 1) and SAN units (4 of Fig. 1) without involvement of the control management station (3 of Fig. 1). Further, the SAN units (4 of Fig. 1) can be dynamically added without interrupting normal data accessing from hosts (1 of Fig. 1) and they are controlled, monitored, and managed by control management station (3 of Fig. 1) through management console (10 of Fig. 1). The control management station (3 of Fig. 1) may also accept storage volume/partition requests from each hosts (1 of Fig. 1), and assign the matched volumes/partitions of SAN units (4 of Fig. 1) to these hosts. Therefore, each hosts (1 of Fig. 1) could directly access the right volumes/partitions of assigned SAN units without goes through control management station again.

This invention will become understood with reference to the following description, claims, and accompanying figures.

Description of Drawings

Fig. 1: Shows an example of simplified block diagram of IP based out-band accessed distributed virtual SAN infrastructure, which includes:

a) **Hosts (1):**

It contains service software modules (9 of Fig. 1). The service software modules (9) communicate with control management software module (7) of control management station (3) to get storage information on a specific IP SAN unit (4). It also communicate with service software modules (6) of IP SAN unit (4) to get block data from SAN units (4 of Fig. 1). The service software modules (9) can be implemented with any suitable programming languages such as C,C++, Java or others and can use any suitable protocols such as IP based or non-IP based or other protocols.

The host could be any system such as a server, desktop or laptop PC, etc., which needs to access block data storage. The spare host (12 of Fig. 1) represents a part of recovery scheme could be implemented in CCDSVM environment.

b) **Network infrastructure (2):**

It represents any kind of communication link, which could be either a department LAN, corporate intranet, Internet infrastructure or others. It consists switches, routers, gateways, cables (Ethernet, optical Fibre, and others), wireless communication media, or others. The network infrastructure provides data path between hosts (1), distribute control management station (3), and SAN Units (4). The network infrastructure also includes software infrastructure such as DNS or DHCP or others to help each systems on the net to find the target address for sending or receiving data within a network domain or in a cross domain environment.

To simplify the discussion, when describing send a message from a system A to a system B, it will simply implied that the DNS or other Internet address identification mechanism is used. In addition, the message is send from source system A to target system B via communication link of this network infrastructure.

c) **Control management station (3):**

It includes distributing control management software modules (7) and console support software modules (8). To support web-based console, it must have web server software (15).

The distribute control management software modules (7) communicate with service modules (6) of IP SAN units (4) to get storage information for constructing a virtual SAN storage pool (11), to monitor IP SAN unit, and to perform various system operations, which includes storage configuration and partitioning etc. It also communicates with service software modules (9) of host (1) for distributing storage volumes to each hosts (1). The distribute control management software modules (7) can be implemented with any suitable programming languages such as C, C++, Java, XML etc. The communication protocols used by distribute control management software (7) between control management station (3) and IP SAN units (4) could be any suitable IP based protocols. The communication between control management station and hosts (1) can be any suitable IP base or non-IP based or other protocols.

The console support software modules (8) get information of the IP SAN units (4) from distributed control management software modules (7) through inter-process communicate mechanism. It further provides these information to web server software (15) through inter-process communication mechanism. The console support software modules (8)

can be implemented with any suitable programming languages such as C, C++, Java, XML etc.

The web server software (15) communicate with management console software (10) on console host (14) through web protocol such as HTTP to provide end user centralized storage management capability for entire distributed virtual SAN infrastructure. The web server software (15) could be an existing commercial software or other proprietary software.

To simplify discussion, the communication path mentioned above will be simply referred as console support software modules (8) communicate (send/receive) with management console (10) on console host (14) without further mentioning the role of web server software (15) on control management station.

In addition, to support non-web based console, there is no needs of web server software (15) on control management station. In this case, the console support software modules (8) could communicate with management console software (10) with a suitable protocol other than web protocol such as HTTP.

d) IP SAN Units (4) and Virtual Storage Pool (11)

The IP SAN contains storage media, storage communication and control media. The storage hardware media of each IP SAN unit might be configured into one or more logical volumes and each volume might has several partitions (Fig. 5).

The IP SAN unit also contains block data service and other service software modules (6), which can communicate with distribute control management station (3) to provide storage information and perform

storage operations. The service software modules (6) also communicates with service software modules (9) of hosts (1) to provide block data service for host. The service software modules (6) can be implemented with any suitable programming languages such as C, C++, Java, etc and the communication protocols used by service software modules (6) can be any suitable IP based protocol.

Multiple IP SAN units are organized and formed a virtual storage pool (11) by control management station (3) in this invention. The virtual storage pool may contain information of each IP SAN unit's IP address, the storage volumes of the block data, their address and sizes etc from each IP SAN units.

The spare IP SAN unit (13 of Fig. 1) represents part of recovery scheme used in central controlled distributed scalable virtual machine environment.

e) **Fibre Channel to IP Gateway (5):**
It translates between Fibre Channel based protocol and IP based protocol so that Fibre Channel based SAN unit will appears as if IP based SAN unit to the rest of world (Fig. 1).

f) **Fibre Channel SAN Unit (6):**
Similar to IP SAN unit except it uses Fibre Channel storage control and connection media and it uses Fibre Channel protocol to communicate with parties. In addition, Fibre Channel SAN unit's (6 of Fig. 2) will appears as an IP based SAN unit to this distributed virtual SAN once it connects to a Fibre Channel to IP gateway (5 of Fig.2). Therefore, to simplify the discussion, it will be treated same as IP SAN unit in all of following discussion without additional comments.

g) **Management Console (10):**

The management console on console host (14), which has been described in pending patent of "**Concurrent Web Based Multi-Task Support for Control Management System**" by the same author. It could be a commercial or a proprietary Web browser, which is able to communicate with web server software (15) on control management station (3) through web protocol such as HTTP. The Web browser could be implemented with any suitable programming languages such as C, C++, Java, XML etc. In addition, the management console software module (10) could be a networked software module other than a web browser. In this case, any other suitable network protocol can be used instead of using web protocol such as HTTP. All of these have been mentioned in section c) above.

To simplify the discussion, the communication path between management console (10) on console host (14) and the console support software modules (8) on control management station (3) will not further mention the role of web server software module (15) in this invention.

From management console (10), multiple concurrent system operations and tasks can be performed for entire distributed virtual SAN infrastructure. There are may be one or more management consoles of distributed virtual SAN infrastructure anywhere on the net.

Fig. 2: This figure is a portion of Fig. 1. It represents the actual virtual SAN. The multiple SAN unit forms a virtual Storage pool (11). The virtual storage pool may contain information of each IP SAN unit's IP address, the storage volumes and their sizes etc.

Fig. 3: This diagram shows a protocol of virtual SAN automatic configuration and building as well as shutdown. The packet format used with this protocol is described in Fig. 4.

Fig. 4: This Diagram shows the message format, which used by "Virtual SAN Automatic Configuration Protocol" for sending and receiving a packet.

Fig. 5: This Fig. Shows the storage in an IP SAN unit, which may be further divided into multiple volumes and each volume may be further divided into multiple partitions. The volume refers to a logical storage unit in this discussion and it might contain multiple piece of storage space from multiple storage hardware media.

Fig. 6: This figure actually is a simplified and a portion of Fig. 1, which shows hypothetical example of how the Storage Volume of IP SAN units can being accessed by hosts. Where each IP SAN units are portion of virtual storage pool (11 of Fig. 2) and each hosts are those same as presented in Fig. 1.

Fig. 7: The Direct Attached Storage System.

Fig. 8: In-Band Accessed Virtual SAN.

This Figure shows another type of virtual SAN, wherein, the actual storage data path from hosts to IP SAN units has to go through control management station.

Fig. 9: A Simplified Diagram of Central Controlled Distributed Scalable Virtual Machine and referred as CCDSVM for brief. With this invention, the systems in a CCDSVM can be flexibly organized into multiple different service pools according to their functionality. For example, multiple IP SAN units can form a virtual SAN storage pool. The hosts of CCDSVM

could form other service pools to provide services other than storage service such as video service, security monitor services, and all other services provided on Web and on net etc.

Fig.10: A Simplified Diagram of Disaster Recovery Scheme of Distributed Virtual SAN Infrastructure, which consists one virtual storage pool of multiple IP SAN units and one service pool of multiple hosts. It assumes that host 1 accesses IP SAN units 1 and 2 while host 3 accesses IP SAN units 4 and 5. It also assumes that IP SAN unit 1 and 2 are mirrored so that they have kept the same copy of data for host 1. The same to be true for IP SAN unit 4 and 5 with host 3. In addition, it assumes that IP SAN unit 3 is a spare unit and the host 2 is a spare host.

Detailed Description of the Invention

1: **Distributed Virtual SAN Infrastructure:**

Fig. 1 Shows a simplified diagram of a distributed virtual SAN infrastructure according to this present invention. With this infrastructure, the distributed virtual SAN storage pool (11 of Fig. 1) comprises one or more SAN units (4 of Fig. 1), which connected to a distribute control management station (3 of Fig. 1) and can be accessed by one or more hosts (1 of Fig. 1) via network infrastructure (2 of Fig. 1). The entire distributed virtual SAN infrastructure can be operated through management console (10 of Fig. 1).

Virtual Storage Pool Auto Building and Initiating:

The virtual storage volume pool (11 of Fig. 1) of distributed virtual SAN infrastructure (Fig. 1) can be initiated and updated when each IP SAN units (4 of Fig. 1) being booted and brought to online, and can be updated when each IP SAN units being shutdown. The Fig. 3 shows the distributed Virtual SAN Automatic Configuration Protocol, which leads to the success of constructing the

virtual storage pool (11 of Fig. 1) of distributed virtual SAN infrastructure (Fig. 1) according to this invention. The following steps have described the automatic building sequence of storage volume pool of the virtual SAN based on this protocol (Fig. 3). The protocol described bellow could be IP based protocol such as SNMP, or a much simple UDP protocol (Fig. 4), any other suitable protocols.

a) When any of IP SAN unit (4 of Fig. 1) such as unit (n) brought up to online, its SAN service modules (6 of Fig. 2) sent out a "SAN unit (n) startup" packet (Fig. 4) to distribute control management station (3 of Fig. 1). This message could be a simple user defined UDP packet (Fig. 4) with message type of system up. This message also could be a SNMP trap of cold start packet, or link up packet (4 of Fig. 1) or other short packet/message of any suitable IP protocols.

b) When distribute control management modules (7 of Fig. 1) of distribute control management station (3 of Fig. 1) receives IP SAN unit (n)'s message, it stores the IP SAN unit (n)'s information.

c) After storing information of the IP SAN unit, the control management modules (7 of Fig. 1) on distribute control management station (3 of Fig. 1) sends back a "need SAN unit (n)'s storage info" packet to IP SAN unit (n) (4 of Fig. 1).

d) When SAN service modules (6 of Fig. 1) on IP SAN unit (n) (4 of Fig. 1) received packet of "need SAN unit (n)'s storage info", it gets storage information on IP SAN unit (n) (4 of Fig. 1), which includes the number of storage volumes, each volume's start address (logical block address, LBA), length, and the end address (logical block address, LBA). The SAN service modules (6 of Fig. 1) then send back a packet of "unit (n) storage info", which includes all information obtained to control management station (3 of Fig. 1).

e) After receiving "unit (n) storage info" packet from IP SAN unit (n) (4 of Fig. 1), the distribute control management modules (7 of Fig. 1) on distribute control management station (3 of Fig. 1) updates its stored information of virtual storage pool (11 of Fig. 1) with corresponding storage information of IP SAN unit (n) from packet.

f) When any IP SAN unit (n) shutting down, the service module (6 of Fig. 1) of IP SAN unit (n) (4 of Fig. 1) sends "Unit (n) shutdown" to distribute control management station (3 of Fig. 1). This shutdown message could be an SNMP trap of link down, or a much simple UDP packet (Fig. 4) with message type of system down, or other short packet based on some other protocols.

g) After received "unit (n) shutdown" packet from IP SAN unit (n) (4 of Fig. 1), the distribute control management modules (7 of fig. 1) on distribute control management station (3 of Fig. 1) updates information of the virtual storage pool (11 of Fig. 1) that specific for IP SAN unit (n) (4 of Fig. 1).

Distributing Storage Volumes in Pool for Hosts Accessing:

After one or more IP SAN units (4 of Fig. 1) are brought into online, the control management station (3 of Fig.1) has owned information of storage volumes and networking for all IP SAN unit (4 of Fig. 1) in the virtual storage pool (11 of Fig. 1). Therefore, the control management station (3 of Fig. 1) is able to distributing storage volumes to hosts (1 of Fig. 1) in several steps.

For example, First, the host 1 (1 of Fig. 1) sends request to control management station (3 of Fig. 1), such as needs 80 GB of storage. Second, the control management station (3 of Fig. 1) stores host 1 information and search for availability of 80 GB of storage volume. It found the volume 2 on IP SAN unit M (Fig. 6). Third, the control management station (3 of Fig. 1) sends the requested information of host 1 to IP SAN unit M (Fig. 6), which include the IP address of host 1, the requested storage size. The control management station (3 of Fig.1)

also sends storage volume information of IP SAN unit M to host 1(1 of Fig.1), which includes the IP address of IP SAN unit M, the volume number and the size, the volume's starting and ending logical address block (LBA) etc. Therefore, all parties of three keep the same storage volume assignment information in sync. Fourth, once the host 1 (1 of Fig.1) and IP SAN unit M (Fig. 6) get each other's information, the host (1 of Fig. 1) can directly and independently access the volume 2 on IP SAN unit M right way with respect of security checking by IP SAN unit M.

Alternatively, These steps may also be semi-automatically setup with assisting of system operations performed from management console (10 of Fig. 1). For example, first an administrator could setup volume 2 of IP SAN unit M (Fig. 6) to be exclusively accessed by host 1 (1 of Fig. 1) as long as he acknowledges that host 1 needs such size of storage volume. Second, the administrator also can setup the host 1 with all information needed to access volume 2 of IP SAN unit M (Fig. 6). Finally, the host 1 (1 of Fig. 1) can access volume 2 of IP SAN unit M (Fig. 6) directly without goes through control management station (3 of Fig. 1).

Dynamic Capacity Expanding:

After distributed virtual SAN storage pool (11 of Fig.1) initiated, the host (1 of Fig. 1) will be able to access the volume on assigned IP SAN unit (4 of Fig. 1) in the pool (11 of Fig. 1) directly without the control management stations' involvement (3 of Fig. 1). This will allow the storage pool (11 of Fig. 1) of this distributed virtual SAN infrastructure (Fig. 1) to continue expanding without effect any hosts (1 of Fig. 1) to continue accessing the storage volumes on assigned IP SAN units (4 of Fig. 1) in the pool. As a result, this guarantees that the distributed virtual SAN storage pool (11 of Fig. 2) can be dynamically expanded without interrupting any normal storage operations and accessing of entire distributed virtual SAN storage pool (11 of Fig. 2).

Scalability:

Once the distributed virtual SAN storage pool (11 of Fig. 1) being constructed, each hosts (1 of Fig. 1) can access one or more IP SAN units (4 of Fig. 1) in the storage pool (11 of Fig. 1) of distributed virtual SAN infrastructure (Fig. 1) whenever it requested. For example, host 1 (Fig. 6) can access IP SAN unit 1, unit 2, and unit M (Fig. 6) after it requested and granted by control management station (3 of Fig. 1). This effectively provides scalable storage system for each hosts (1 of Fig. 1) within distributed virtual SAN infrastructure (Fig. 1) of this invention. Further, the distributed virtual SAN infrastructure (Fig. 1) provides far better scalability than the in-band accessed virtual SAN (Fig. 8), wherein, the scalability of in-band accessed virtual SAN were severely limited by the bottlenecked control management station (Fig. 8).

Storage Sharing:

Once the distributed virtual SAN storage pool (11 of Fig.1) being constructed, each IP SAN units (4 of Fig.1) in the pool of distributed virtual SAN infrastructure (Fig. 1) may be hold multiple storage volumes in the form of block data for one or more hosts (1 of Fig.1) accessing. Therefore, it will allow multiple hosts (1 of Fig. 1) to share an IP SAN unit (4 of Fig. 1) by granting and assigning each host to exclusively access particular volumes on that IP SAN unit (4 of Fig. 1). The Fig. 6 demonstrates such a storage sharing, wherein, IP SAN unit 2 of Fig. 6 has three volumes, which named volume 1, volume 2, and volume 3. The block data service modules (6 of Fig. 1) on IP SAN unit 2 of Fig. 6 can arrange share volume 1 with host1 and shares volume 2 with host 2 exclusively.

Performance:

With in-band accessed virtual SAN (Fig. 8), the control management station could be a performance bottleneck. With distributed virtual SAN of this invention each hosts (1 of Fig. 1) can directly and independently accessing any IP SAN unit (4 of Fig. 1). Therefore, the performance of storage accessing for each hosts will not be effected and can match the performance of direct attached storage system (Fig.

7) when the high speed network connecting media is deployed in distributed virtual SAN infrastructure (Fig. 1).

Centralized Management of Distributed Virtual SAN:

The storage management console on a console host (10 of Fig.1) can communicate with console support software module (8 of Fig. 1) on control management station (3 of Fig. 1) and to further get information of all IP SAN units (4) from control management modules (7 of Fig. 1) of control management station (3 of Fig. 1). Therefore, it can provides centralized management functionality for entire distributed virtual SAN storage pool (11 of Fig. 1), hosts (1 of Fig. 1), and the control management station itself (3 of Fig. 1). With multiple concurrent tasks supporting in console support software module (8 of Fig.1) of control management station (3 of Fig. 1), the storage management support console (10 of Fig.1) can provide a full range of system operations and tasks. In addition, multiple system tasks and operations can be run concurrently throughout the entire distributed virtual SAN and hosts. These management tasks includes storage configuration, storage volume allocation and assignment, storage partition and repartitioning, storage, network, and other resource usage and activities monitoring, etc..

Disaster Recoverability:

The use of DNS or may be other IP address identification mechanism helps this distributed virtual SAN infrastructure overcome the geometric region limitation and work well in either cross network domains environment or in a single network domain environment. Therefore, any IP SAN unit or host as well as control management station could be anywhere on corporate Intranet, on department LAN, or on Internet. As a result, it is possible to have a disaster recoverability plan goes beyond 100 miles long vs traditional 10 kilometer limitation.

In addition, the disaster recovery plan of distributed virtual SAN infrastructure can be flexibly implemented as showing in Fig. 10. With this recovery plan, the host 1 or 3 (1 of Fig. 10) can continue to operate whenever one of its mirrored IP SAN units failed (3 of Fig. 10). Also, the spare IP SAN unit can be used to quickly replace the failed IP SAN unit whenever there is needs. On the other hand, the hosts (1 of Fig. 10) also can be organized into a service pool such as for distributing video service, distributed database pool, distributed security monitor services, and all other services provided on net and on Web. Therefore, whenever host 1 or 3 failed, the spare host can quickly take over their assigned IP SAN storage and replace them to continue provide service to the end user.